DOCUMENT RESUME

ED 401 307 TM 025 868

AUTHOR McBride, Bethe; Carifio, James
TITLE Empirical Results of Using an Analytic versus Holistic Scoring Method To Score Geometric Proofs: Linking and Assessing Greeno, Bloom, and van Hiele Views of Student Abilities To Do Proof.
PUB DATE Apr 95
NOTE 56p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 18-22, 1995).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Analysis of Variance; Cognitive Processes; Correlation; Criteria; Evaluators; *Geometry; High Schools; *High School Students; *Holistic Approach; Mathematical Aptitude; *Mathematics Achievement; Reliability; *Scoring
IDENTIFIERS *Analytic Scoring; *Van Hiele Levels

ABSTRACT

This study sought to establish the benefits of an analytic scoring procedure for assessing student performance in doing geometry proofs. Using the cognitive behavior theories of B. Bloom and the theories of J. G. Greeno about geometric knowledge, five criteria were established for assessing performance in proof. After a training session, 3 judges rated the performance of 241 students using the scale on a proof test used in previous research or a newly developed proof test. Student proof scores were established through averaging the judges' ratings. The scores were compared with the scores of other tests intended to measure a student's ability to do proof. Repeated measures of analysis of variance were used to determine the interrater reliabilities for all the ratings, which were consistently high. Cronbach's alpha was used to estimate the internal consistencies of test scores resulting from the new scoring method, which were very high. A step-down analysis of variance for three methods of scoring proof, by proof test form, was conducted to investigate the analytical sensitivity of each method, revealing the analytic scoring method to be much more statistically sensitive. A table of correlations for van Hiele levels and the new assessment criteria was generated, and a convergent/discriminant analysis of correlations revealed that the scoring criteria were not well aligned with van Hiele theory's of geometric thinking. An appendix presents the Revised van Hiele Test. (Contains 6 tables and 16 references.) (Author/SLD)

Empirical Results of Using an Analytic Versus Holistic Scoring Method to Score Geometric Proofs: Linking and Assessing Greeno, Bloom, and van Hiele Views of Student Abilities to Do Proof.

Bethe McBride & James Carifio

University of Massachusetts at Lowell

Running head: Analytic versus Holistic Methods

Abstract

This study sought to establish the benefits of an analytic scoring procedure for assessing student performance in doing geometry proofs based on the theories of Greeno and Bloom. Five criteria were established for assessing performance in proof. After a training session, three judges rated the performance of 241 students using the scale on a proof test used in previous research or a new proof test we developed.

Student proof scores were established through the averaging of the judges' ratings. To establish the validity of scores generated by this type of assessment, the scores were compared with the scores of other tests intended to measure a student's ability to do proof. Repeated measures ANOVAs were used to determine the interrater reliabilities for all the ratings, which were consistently high. Cronbach alpha was used to estimate the internal consistencies of test scores resulting from the new scoring method, which were very high. A step-down analysis of variance for three methods of scoring proof, by proof test form, was conducted to investigate the analytical sensitivity of each method, which revealed the analytic scoring method to be much more statistically sensitive. A table of correlations for van Hiele levels and the new assessment criteria was generated and a convergent/discriminant analysis of correlations revealed that the scoring criteria, a combination of Bloom's taxonomy with Greeno's geometric knowledge areas, were not well aligned with the van Hiele theory.

The purpose of our research was to explore the possible benefits of a different method of scoring geometry proofs. The different method was an analytic, as opposed to dichotomous or holistic, scoring approach. The analytic scoring scale we constructed consisted of five major criteria. The major criteria were broken into subcriteria applied to specific aspects of performance in geometric proof, thereby giving a very detailed description of a subject's proof development behavior as well as the degree of correctness of the proof (see below for details).

This article reports the results of our research relative to investigating the benefits of using an analytic rating scale to score geometric proofs. These hypothesized benefits were: (a) increased internal consistency of the scoring of proofs, (b) a more normal distribution of scores, (c) more detailed and precise descriptions of proof development behavior, and (d) greater validity of scores through the use of scaleable subtests. The background theory related to the development of the scaling method will be presented prior to giving the research results.

Background: The Analytic Scoring Method

After developing a new geometry proof test and a new geometry problem-solving test, due to limitations in those currently available, we considered the question and problem of how to score these new tests. At first, a very simple *right-or-wrong* traditional scoring method was considered; namely, a method where a team of three judges would determine independently whether the subject had responded correctly to a proof and then the consensus of their opinions would be used to score the problem or proof either right or wrong. This simple method of scoring, however, would not be very effective in validating the two tests we developed or in doing analyses of results from these two tests. Such a dichotomous method would result in scores of 0 to 6 for the New

Proof Test (NPT), since there were 6 proofs, and 0 to 5 for the Geometry Problem Solving Test (PST). We wanted a method that would be more statistically sensitive and descriptive, and one that did not truncate variances so severely. An analytic scoring method, therefore, was developed for use. The analytic method was based on a method presented by Charles, Lester, and O'Daffer (1987, p. 30), which focuses on measuring degrees of correctness in proofs and mathematical problem solving. To the best of our knowledge, this performance philosophy and scoring approach has not been used in research on geometry and proofs although it has been advocated by several prominent theorists (e.g., Polya, 1957, and Greeno, 1978).

Next we considered the issue of whether the analytic scoring method could be used for both the proof and problem-solving tests. We believed the same scoring rubric could be used if the *same* criteria were used for analyzing performance on both tests and a conceptual link between problem solving and proof generating behaviors could be established (see below). The analytic scoring method, therefore, was designed to be used for both the PST and the NPT. The remainder of this background section discusses proof as a type of problem, the typical types of student proof performance seen in secondary-school geometry, the cognitive abilities necessary for successful proofs and problem solving, and the details and mechanics of the analytic scoring method used on this research.

Proof as a Type of Problem

Initially, one must distinguish between the possible interpretations of the word "proof." In one context a proof can be a finished argument; as in "the professor presented his proof." The use of the term "proof" here implies the argument is completed and ready for presentation. The sense in which a proof can be regarded as a problem

is when the term refers to a task that a student is asked to complete, as when a teacher says "do the proofs on page 421 for homework."

Further, one needs to understand the distinction between a *problem* and an *exercise*. A problem can be defined as a task for which the subject does not have an immediate procedure for solution (Lester, 1980). An exercise is a task for which the subject knows a method of solution. Newell and Simon further reinforced the definition of problem when they wrote: "A person is confronted with a problem when he wants something and does not know immediately what series of actions he can perform to get it" (1972, p. 72). Glover, Ronning, and Bruning have stated: "A problem exists when an obstacle separates our present state from some desired state" (1990, p. 149). A proof fits all these definitions of a problem and definitions supplied by others such as Polya (1957) and Greeno (1978).

A true geometric proof is an original argument provided by its author. Though the statements and reasons provided by the student may be the same as those of another author, a true proof is not memorized and then written or recited. In a true proof, the student is provided information and asked to show that that information leads to a specified conclusion. The obstacle, as Glover et al. (1990) referred to it, is finding the right chain of deductions and presenting them coherently. To write a true proof is to fill the logical gaps between the *givens* and the *prove*. If the student simply lists or recites a series of statements and reasons, he has not accomplished a successful proof. He has mimicked someone else's proof.

In some instances, a student may be asked to prove a well-known theorem. This can still present a problem to those who have not memorized a proof of that theorem. In research, it would be difficult to tell if a student worked out a proof on her own or if she memorized

the proof. Using proofs that are not typical theorems in a text, therefore, would be critical in an instrument designed to measure proof ability. In such instances, the ability to author a competent proof based on algorithmic behaviors is low probabilistically. A set of specific instructions for doing categories of proofs cannot be provided by a teacher. Because a student must use heuristics and judgment (both strategic and tactical) to arrive at a true proof, a geometric proof's solution path is not immediately evident with such proofs.

Heuristics come in when the student begins to make deductions based on the information and the problem figure. Not all the conclusions the student makes necessarily contribute to the argument. Students can believe that irrelevant deductions are important. Sometimes, students attempt a proof using the wrong strategy, like using a triangle congruence plan when that strategy cannot be successful. Other times, students infer spatial relationships from a diagram that negate the need for giving a proof. In a true proof, the series of statements and reasons must be discovered. Thus, it appears that a true geometric proof is a type of problem. A proof is a problem because it is a task that is slightly "fuzzy" in several ways rather than precise with only one solution and solution path.

Several researchers have used proofs in their examples of problems. Among these authors are Newell and Simon (1972, p. 73) and Greeno (cited in Glaser, 1978, pp. 50-52). Newell and Simon use the proof task as an example of a well-defined problem because the desired result is so specific and any proposed solution is testable. It would seem then that a proof is a type of problem, and as such any broadly inclusive scheme designed to measure the ability to solve a problem would also most likely apply to the ability to write an original proof.

Typical Student Behavior

How do students typically perform when confronted with a proof and

how does a teacher assess the student's performance? Answering these questions is critical to developing an instrument to measure the ability to do proof. Therefore, let us consider how students typically perform. Based on more than twenty years of correcting geometric proofs, many conversations with colleagues who have corrected proofs, and conversations with geometry students, the following characterization of student performance is proffered.

Typical high school students fall into several very broad categories when it comes to solving a proof problem. For any given problem, some of the possible student performance categories might include:

1. Unschooled: Those who have no idea what to do.

2. Novices: Those who understand what they are expected to attempt but are confused.

3. Intermediates: Those who show an understanding of the process but appear to be missing knowledge of a key concept that would enable the necessary deductions.

4. Competents: Those who appear able to solve most problems or write most proofs.

5. Experts: Those who can write eloquent proofs or problem solutions.

The above topology is helpful in a number of ways. Therefore, let us review some of the behaviors of members of the above categories.

The Unschooled, who have no idea what to do, may write down the given and attempt a statement or two mimicking another proof performance they have witnessed. The Unschooled may have some knowledge of facts or propositions; however, they cannot make inferences about abstractions and cannot make plans to solve the problem. The Unschooled frequently do not comprehend the problem space.

The Novices do show indications of comprehension. The Novices often say that they understand a theorem or definition, but cannot apply

it. Novices do not appear to discriminate well the relevant from the salient. The Novices are easily misled. They will use notation correctly and can match terms to their definitions. In a proof, the Novices always write down what is given, they usually pick something else out they suspect to be true based on the diagram (e.g., if it looks like a rectangle, they assume the angles are right angles) and they usually put the correct statement on the last line. Sometimes the Novices make good deductions. The Novices appear to have favorite reasons like "SAS" or "Reflexive" and they usually try to work them into the proof somewhere. These preferences affect the strategies of the Novices and they often try to use a plan that is unworkable. Novices will make characteristics of a shape part of the given instead of simply classifying the figure.

The Intermediates have the "big" picture regarding proof but they stumble over their statements and reasons because they have gaps in their knowledge base. The Intermediates might abbreviate too much, leaving out essential phrases. The Intermediates may leave the given out of the proof. They may make up their own theorems (to fill the gaps) and invent reasons. Intermediates frequently believe their missing steps are obvious and need not be written.

The Competents will be able to correctly do most proofs from a typical high school text. The Competents will remember the essential theorems and definitions as well as postulates. The Competents generally do not write any unnecessary steps although they may not take the most efficient path to a solution. The Competents exhibit good problem-solving skills. The Competents do not put more information in the given than is necessary.

The Experts are able to solve problems in a minimum amount of time and without wasted efforts. The Experts will quickly seize upon a strategy that leads to the solution with an economy of steps and

simplicity of explanation. They make effective use of previous theorems, definitions, and postulates. Experts are able to produce "elegant" proofs.

Although useful in many ways, these five broad categories alone do not provide adequate criteria for research or assessment. One objective of this research was to assess or categorize student performance optimally when it comes to proof or problem solving. The better the system of assessment, the better the range of possible scores or categories, which, in turn, will improve the statistical analyses. A closer look at some of the essential behaviors affecting the student's ability to do a proof, therefore, will help guide the process of finding an assessment method.

Necessary Cognitive Resources

An analytic scoring method requires an in depth assessment of the types of knowledge and cognitive behaviors that are associated with success in doing proofs. Doing a proof requires attention, perception, memory-search, heuristics, and logic. Memory search is critical to success in any problem-solving situation. The student must be able to access relevant information to make the links between the knowns and the objective. Having the right kinds of knowledge can make the difference between success and failure in doing geometry proofs.

Greeno (1978) has given a reasonable analysis of what knowledge is needed to solve geometry problems and hence to do proofs. He refers to four requisite types of knowledge: (1) general knowledge such as the definitions of terms and the meanings or markings, (2) visual knowledge of angle formations such as vertical angles or patterns that imply special relationships between geometric figures like triangles, (3) knowledge of inferential propositions (e.g., postulates or theorems), and (4) knowledge of strategies or possible plans for reaching the objective such as planning to use SAS to prove triangles congruent or

the Pythagorean theorem to find a missing side. Of the four types of necessary knowledge, the first three are declarative and the fourth is procedural. Further, a successful proof or problem solution can depend on knowledge and appropriate use of notation, units, and symbols.

In addition to the knowledge resources the student must possess to achieve competence at proof, the student must exhibit certain thought processes. Bloom (1956) proposed a taxonomy of cognitive behaviors that arranges thought processes into a hierarchy. At the lower end are the recall processes and at the upper end are reflective processes. Bloom's taxonomy can be used as a guide in assessing a student's progress in problem solving. Using Bloom's taxonomy combined with the knowledge bases suggested by Greeno as a guide, it is possible to outline the spectrum of cognitive behaviors that are exhibited in the process of doing a proof.

Bloom's taxonomy places cognitive behaviors into a hierarchical sequence. The later behaviors are considered higher level cognitive functions and build on those that come before. Bloom's hierarchy has been shown by factor analysis to resemble a letter Y with *knowledge*, *comprehension* and *application* forming the trunk, *evaluation* and *analysis* being on different limbs, and *synthesis* an extension of analysis.

Using Bloom's taxonomy of cognitive behaviors as a guide, the essential behaviors necessary for doing proof and solving geometric problems are given below:

1. The student requires knowledge of terms, patterns, symbols, definitions, properties, propositions, strategies, syllogisms. A student gives evidence of possessing the required knowledge by: (a) drawing a diagram that can be used to make the proof; (b) using the terms mentioned in the problem in a meaningful context; (c) mentioning formations or patterns that occur in the problem; (d) using symbols, syntax, and notation correctly; (e) referring to relevant postulates, theorems, or definitions; (f) making at least one correct deduction; and (g) assembling a series of statements and reasons that shows the intent of accomplishing some subgoal, if not the ultimate goal.

2. The student must comprehend the problem space as related to the proof. What is known? What exactly needs to be shown or done?

Students exhibit their comprehension of the problem space when: (a) the relevant "given" are used in the body of the proof or the calculations; and (b) the last line of the proof is the statement required in the proof accompanied by a plausible reason, albeit not substantiated; or the problem's answer is given in terms of the proper units or phrases.

3. The student needs to be able to apply deductive logic to knowns and previous deductions and to apply the relevant theorems or definitions. The student also needs to apply the proper notation and syntax. A student exhibits this behavior when his work: (a) includes at least one valid use of a rule of logic or a theorem; and (b) when his notation is correct.

4. The student must be able to analyze the problem. How do the "givens" affect the outcome? What is relevant? What can be ignored? What else needs to be known (deduced)? A student exhibits this behavior when: (a) only necessary statements or formulae are included in her work; (b) relevant information is deduced from the givens; and (c) relevant statements are deduced and written, albeit the reasons supplied for them may be incorrect.

5. The student needs to be able to synthesize the information and to be able to make a *plan* for closing the gap between what is known and what needs to be proven. The bridge between knowns and the objective will be constructed from deduced facts. What type of strategy (for example, SAS or CPCTC) should be employed to reach the ultimate goal of deducing the requested proposition or to answer the question? Generally the last lines of a proof or the initial equation of a problem exhibit the main solution strategy intended by the author. Although the student may have missed some pieces (i.e., not known why two relevant parts were congruent), it can be inferred that the student attempted to employ a viable strategy when: (a) the penultimate line is relevant or the pivotal equation is appropriate, and (b) the statement on the last line is the objective of the problem, accompanied by a logical reason, or the solution and units are correct.

6. The student needs to be able to evaluate his proof. Is the evidence I have given relevant? Are my deductions worthwhile in the context of this problem? Have I accomplished the proof? Is it the best possible proof or was there a better way to do it?

The student's ability to evaluate in the Bloom sense (i.e., to assess or detect the logical consistencies or inconsistencies of arguments) is assessed by the overall correctness of the proof itself. Evidence of the evaluation process is found in corrections to text. Evaluation is evident when: (a) a student appears to have abandoned an irrelevant statement or equation by erasure or scratch-out, or (b) the student adds statements and renumbers lines.

By comparing typical student performance with the necessary cognitive processes mentioned earlier, we determined criteria for assessing performance that could be observed in a student's written proof or problem. The criteria do not correspond one-to-one with Bloom's six

categories of cognitive function nor Greeno's areas of geometric knowledge (see below). The lack of the one-to-one correspondence is due in part to the fact that some cognitive behaviors mask or combine with others. For example, it is difficult to separate the categories *analysis* and *synthesis* just by reading a proof. For example, the plan a student embarks upon depends on his analysis. Further, knowledge of relevant theorems, properties, or rules of logical inference is frequently only evident when the knowledge is applied. Therefore, the criteria of the analytic scoring method were geared to combinations of the cognitive processes suggested by Bloom and the knowledge areas suggested by Greeno.

The Analytic Scoring Criteria

With both Greeno's and Bloom's theories as a guide and with many years of experience in correcting student proofs, the author proposed these five criteria for assessing student performance on written geometric proofs and problems:

Criterion I (Knowledge): The student appears to have knowledge of the terms, notation, and diagram markings of the problem.

Criterion II (Comprehension): The student has determined the known information and the objective of the problem.

Criterion III (Application): The student uses pertinent definitions, postulates, theorems, etc., and key visual formations of the problem to develop the proof or solution.

Criterion IV (Analysis & Synthesis): The student has a strategy or plan to accomplish what he or she intends and knows when an auxiliary part (e.g., segment) is needed.

Criterion V (Evaluation): The student can make valid deductions and present an effective proof/problem-solution.

The above criteria also show the integration of Greeno's theorized knowledge areas. Criterion I relates to general knowledge of terms, markings, and so on. Criterion II relates to visual knowledge since frequently understanding the problem includes knowledge of visual

formations. Criterion III can be linked to Greeno's knowledge of inferential propositions, and Criterion IV corresponds to Greeno's knowledge of strategies.

Each criterion was subdivided by having the judge apply the criterion to particulars. For example, on the proof analytic scale scoring sheet, Criterion I was assessed relative to knowledge and use of (1) terms, (2) notation, (3) diagram markings, and (4) overall impression of these basic requisites. Further, not all five scoring criteria had the same number of subdivisions. A sample of a proof test analytic rating scale scoring sheet is given in Table 1, which is given in the results section so that the interrater reliability for each judgment may be included in the table presenting this rating scale.

The ratings for each subcategory ranged from 0 to 4 depending on the degree to which the subject's work met the scoring criteria. The definition of these rating codes were:

Code 0: Only negative (wrong) indicators are present (Contra indicated). A rating of 0 is given when the subject's work indicates the criterion has not been met.

Code 1: No indications or indeterminate (Cannot tell). A rating of 1 is given when it cannot be determined if the criterion has been met. Either there is no evidence or the subject's work does not relate to the specific criterion being applied.

Code 2: Only a few positive indicators are present (Slightly). A rating of 2 is given when the subject's work indicates the criterion has been met to a very limited degree. The subject may appear to understand some of the terms or one of the relevant theorems but does not appear to have enough knowledge to make a significant effort to solve the proof or problem.

Code 3: Mostly positive indicators (Mostly). A rating of 3 is given when the subject's work almost completely satisfies the criterion. The subject appears to know all but one of the terms or theorems, or to be very close to a correct solution.

Code 4: Indicators are all positive (All). A rating of 4 is given when the subject's work fully meets the criterion.

The above scoring codes were right on the proof or problem-solving scoring sheet to make it simpler for the rater to use these codes (See Table 1).

Method of Validating Analytic Scoring Method

It is essential to show that the analytic scoring method is preferable to other, less time intensive methods of scoring. For this reason the analytic scoring method was compared to two other methods of scoring. The comparison involves not only the types of scoring methods, but four of the instruments used in the main research.

Instruments Used to Validate Analytic Scoring Method

The four instruments involved in this research include both open-ended and multiple-choice tests. The open-ended (generative) question tests were proof tests. As mentioned previously, a new proof test was developed. A proof test developed by the Cognitive Development and Achievement in Secondary School Geometry (CDASSG) team was also used for the purpose of comparison and as a marker. The CDASSG Proof test chosen was Form 3. The reason Form 3 was chosen was that it had the highest reliability rating of the three forms (Cronbach α = .88) and the items all had mean scores between 1.66 and 2.86 on a 0 to 4 scale. For the main research both proof tests were scored using the analytic scoring method the author developed.

Along with the New Proof Test (NPT) and the CDASSG Proof Test (CPT), two other instruments were also used. One was a test to determine a subject's van Hiele (Teppo, 1991) level of geometric thinking known as the Revised van Hiele Test (RVHT) and the other was a test used to categorize subjects by Piagetian stage known as the Equilibrium in a Balance Test (EBT) developed by Adi (1976). It should be noted that these four instruments formed the core of a convergent-discriminant validity study (Campbell & Fiske, 1959).

The EBT is a 15 item multiple-choice response test. It is broken into three subtests of five items each. Each subtest is associated with a Piagetian operational or cognitive stage, (i.e., concrete, transitional, formal). The items ask the subject to choose a solution that will keep a balance in equilibrium. In previous research, the EBT had a Guttman reproducibility coefficient of .96 and a scalability coefficient of .87 (Adi, 1980).

The Revised Van Hiele Test (RVHT) also uses a multiple-choice response mode and consists of three eight item subtests. Each of the three RVHT subtests corresponds to a van Hiele level of geometric thinking (i.e., Visual, Descriptive, Theoretical). The items all relate to geometric shapes and their possible relationships. The RVHT psychometric findings revealed a Guttman reproducibility coefficient of .97, a scalability coefficient of .80, and a level assigned test-retest correlation of .83.

The EBT and RVHT tests were multiple choice and could be scored by a computer. The proof tests were scored by three trained geometry teachers using the analytic scoring method previously explained. These three judges were trained as previously described.

Both the CPT and the NPT consist of six problems. Each has a problem that requires filling-in missing parts of a proof, a problem asking for the set-up of a proof, and four complete proofs. The CPT has no simple prove-these-triangles-congruent problem nor any problem involving circles. The NPT features a circle problem and a simple triangle congruence problem. Both tests consist of typical high school textbook proofs.

Judges and the Scoring Procedure

Three judges were used to score proofs using the analytic scoring scale given in Table 1. Each judge was a high school geometry teacher, all with at least ten years of experience. All the judges had master's

degrees, two were Ed.D. candidates, and the other a department head.

The judges were trained to do the scoring. After instruction, they discussed the criteria and the written evidence that would correspond to each criterion. Then the judges practiced the method of scoring on identical samples of unrelated student proofs until each was comfortable with the scoring rubric. The rubric was formalized and the judges continued to practice until they agreed on approximately 80% of the ratings they made.

Each analytic scoring criterion was applied to three specific problems. There were two reasons why each criterion was applied to only three proofs or problems. First, not all problems lent themselves to giving evidence of the criterion being applied. Secondly, applying the criterion to only the most suitable items would shorten the labors of the judges. The proofs for each criterion were chosen because they were likely to contrast the behaviors of students for that criterion. We attempted to choose three very suitable proofs for each criterion and score each proof at least twice.

Participants

The 241 high school geometry students took one of two proof tests. One proof test had been used previously in research by the Cognitive Development and Secondary School Geometry (CDASSG) team, which developed it. This test is referred to as the CDASSG Proof Test (CPT). The CPT was a reliable proof test ($\alpha = .88$); however, it covered a more limited number of proof topics (for example, no circles) than the New Proof Test (NPT) developed for this study.

Test Administration and Handling

The tests were administered in late May and early June of 1993. There was a total of 253 subjects from Massachusetts and southern New Hampshire involved in different aspects of the research. Most of the

subjects were from suburban communities; 73 were from an urban high school. There were 220 subjects who took both the EBT and the RVHT. Of the 241 who took a form of proof test, 119 took the NPT and 122 took the CPT.

All tests were administered by a given student's geometry teacher. There were seven classroom teachers from four secondary schools involved in administering the tests. The teachers were supplied with instructions and scripts for administration. The EBT, RVHT, and one of two proof tests (CPT or NPT) were administered on three separate days.

The CPT and the NPT forms had been prepared separately and then shuffled together numerous times before they were handed to the administering teachers so that there would be no controlling the test form that a student received. Each of the four participating schools had roughly the same number of proof forms (CPT and NPT) to distribute. The classroom teachers were told to pass out the proof-test forms themselves to insure the randomness of the distribution of the two forms. As the present research was primarily aimed at determining the performance of an analytic scoring method and convergent and discriminant validity, strict randomization was not as important as in a treatments comparison study.

We made three copies of each subject's proof test. The NPT's were copied in pink and the CPT's in white. Next we generated a series of random numbers to determine the order for correcting the proof tests, by version. Six packets of approximately 40 tests (N = 241) were prepared for each judge consisting of tests following the randomly determined sequence.

Analytic Scoring Method

There was a total of 60 ratings made per subject, per judge. Using the analytic scoring method for each test, the minimum score was 0 and the maximum test score was 240 (60 times 4). The closer a score was

to 0, the more likely the student's work showed little or no ability to solve a problem or do a proof. The closer a student's score came to 240, the more the student's work was at the expert level.

The final scores were essentially ordinal in nature because the gaps between the integer scores would not necessarily indicate the same difference in performance. However, according to Kerlinger (1986, p. 494), a numerical rating scale can approach interval status, primarily as a function of the number of items or ratings made on the scales. As there were 60 five-point ratings on our scale, we treated the aggregate scores from the scale as interval data in this study.

The student's final score was the total of the average of the judges' ratings on each of the 60 measures. That is, the three judges' scores were averaged for each judgment and the total of these averages was the final score used for analyses. Final scores also ranged, therefore, between 0 and 240.

When the two multiple-choice tests (RVHT and EBT) were scanned, the scores of these tests and specified subtests were saved in addition to subjects' actual choices (for example, A, B). For example, on the Revised Van Hiele Test, the scores on subtests for the first-eight items, second-eight items, and third-eight items were recorded in addition to the scores for the full 24 items. The EBT was broken into three subtests by successive groups of five items.

Subdividing the tests allowed us to quickly establish a revised-van-Hiele-level and Piagetian-level score for each subject. The process for assigning a revised-van-Hiele-level was as follows:

1. Establish a cut-off score for mastery at each level (e.g., let cut-off be 6).

2. Assign Level-1 score equal to 1 if the first-eight score exceeds or equals the Level-1 cut-off score, otherwise let the Level-1 score be 0.

3. Assign a Level-2 score equal to 2 if the second-eight score equals or exceeds the Level-2 cut-off score, otherwise let the score be 0.

4. Assign a Level-3 score of 4 if the third-eight score exceeds Level-3 cut-off score, otherwise assign a score of 0.

5. Assign subject Level 3 mastery if all three level scores total seven. Assign subject Level 2 mastery if the level scores total three. Assign subject Level 1 mastery if the level scores total one. Assign student Level 0 mastery if the level scores total zero. Subjects with any other level totals cannot be assigned a mastery level.

Students with level totals other than 0, 1, 3, or 7 had failed to show mastery at a level lower than the highest level they mastered and thus did not satisfy the hierarchical criterion for van Hiele level assignment. A similar approach could be used to sort students into Piagetian stage using the EBT.

On the EBT the subjects were given scores based on their number correct and by a theorized level. The same method was used for the RVHT. Subject's scores on the NPT and CPT were a total of all the averaged ratings done by the three judges.

Comparing Proof Scoring Methods

To make a comparison of scoring methods possible, the two proof tests, NPT and CPT, were scored three different ways. The scoring methods are as follows:

1. Dichotomous: Score each of the six proofs on each subject's test either right or wrong. The "Net" ratings at the right end of a row on the analytic score sheet (see Table 1) were used to determine if a proof was right. A proof was right if the total of the averaged net values for a specific proof was 75% or more than the total possible value. For example, referring to the sample analytic score sheet displayed in Table 1, Item 2 (i.e., Proof 2) appears in Criteria I, III, and V. This means the total possible "Net" points for this proof would have been 12 because the highest average "Net" score for each Criteria was four and there were three occurrences of ratings on Item 2. We used 75% or more as the cutoff for being right. So, if the total averaged "Net" ratings was nine or more for Item 2 the proof was considered right. A one was assigned to proofs deemed right, and zero if deemed wrong, and the subject's score was the total of the one's received. This way of scoring could result in scores for each student ranging from zero to six.

2. Holistic: Use the average of the "Net" ratings for each proof to calculate the final sore. For example, again referring to Item 2 on the sample analytic score sheet, the score for Item 2 (i.e., Proof 2) was the average of the three "Net" ratings found in Criteria I, III, and V. Each proof on the test received an averaged score and these six scores were totaled. In this way each proof received a score from zero

to four and each subject's test resulted in a score ranging from 0 to 24.

3. Analytic: Score each test using the analytic scoring method previously described. The analytic scoring method could result in scores ranging from 0 to 240.

The results presented below compare these three scoring methods in terms of reliabilities, correlations, convergent and discriminate validity, and score distributions at both the total test and subtest levels. Lastly, we considered the potential uses of the analytic scoring criteria as subtests. To this end, we constructed tables of correlations for the five analytic criteria. In these tables we compared the van Hiele theory with an integrated theory including Bloom's taxonomy and Greeno's geometric knowledge areas.

Results

The interrater reliabilities of the analytic scoring method were considered and assessed first. Involving judges introduces the element of "errors of judgment" into the already complex field of possible sources of error. As an effect of having the items scored by judges, the judges become a feature of each item (i.e., proof). Three judges were used in an effort to control the level of this type of error. However, the reliability of the three judges was a concern. Therefore, repeated measures analysis of variance was used to study the judges' patterns of ratings.

The overall agreement or disagreement of the judges' ratings is evident in the separate interrater reliability coefficients (IRC's) for each of the 60 ratings the judges made per subject, and for their rating totals by subject. These IRC's are given in Table 1. Table 1 gives the alphas for each of the 60 ratings made by the three judges over all 241 subjects (it combines both proof tests). The entry in the blank beside "Total" in Table 1 shows the IRC's for the judges' rating totals over the same subjects.

Table 1: *Interrater Reliability Coefficients (IRC's) for the Judges over Both Proof Tests, CPT and NPT (N = 241)*

SCORE SHEET

(Code 0) Only negative indicators (Contra indicated) (Code 3) Mostly positive indicators (Mostly)
(Code 1) No indications or indeterminate (Can't tell) (Code 4) Indicators are all positive (All)
(Code 2) Only a few positive indicators (Slightly)

Criterion I

1) **The student appears to understand the terms, notation, and diagram markings of the problem.**

Apply the above codes to:					Specify	
Item 2—Positive Indicators present (code)	T .88	N .72	M .63	Other		Net .87
Item 4—Positive Indicators present (code)	T .83	N .80	M .59	Other		Net .86
Item 5—Positive Indicators present (code)	T .89	N .79	M .70	Other		Net .87

Criterion II.

2) **The student has determined the known information (semantic or visual) and the objective of the problem.**

Apply this above codes to:					Specify	
Item 3—Positive Indicators present (code)	S .79	V .63	O .84	Other		Net .84
Item 5—Positive Indicators present (code)	S .87	V .73	O .88	Other		Net .91
Item 6—Positive Indicators present (code)	S .80	V .84	O .88	Other		Net .90

Criterion III.

3) **The student exhibits knowledge of pertinent definitions, postulates, theorems, etc., and key visual formations of the problem.**

Apply the above code to:					Specify	
Item 1—Positive Indicators present (code)	D .88	T .91	V .85	Other		Net .92
Item 2—Positive Indicators present (code)	D .87	T .86	V .61	Other		Net .88
Item 4—Positive Indicators present (code)	D .87	T .86	V .87	Other		Net .90

Criterion IV.

4) **The student has a strategy or plan to accomplish what he or she intends and knows when an auxiliary segment is needed.**

Apply the above code to:				Specify	
Item 1—Positive Indicators present (code)	P .92	A .03	Other		Net .92
Item 3—Positive Indicators present (code)	P .89	A .82	Other		Net .89
Item 6—Positive Indicators present (code)	P .85	A .77	Other		Net .86

Criterion V.

5) **The student can make deductions and present an effective proof.**

Apply the above code to:						Specify	
Item 1—Positive Indicators present (code)	D .91	C.90	FD .92	R .88	Other		Net .92
Item 2—Positive Indicators present (code)	D .92	C.85	FD .88	R .89	Other		Net .92
Item 3—Positive Indicators present (code)	D .83	C .81	FD.83	R .80	Other		Net .86
					Total:		.95

As can be seen in Table 1, the IRC' predominantly ranged between .59 and .92. The squares of the IRC's shown in Table 1 can be interpreted as approximations of the percent of times the judges agreed on a rating. Overall, the mean IRC for the three judges was a very impressive .83, which suggests an estimated interrater agreement ratio of 69%. One sees from the typically high IRC's in Table 1 that the judges were very similar in their ratings. The IRC in the "Total" space at the bottom of Table 1, IRC = .95, indicates how very similar the

judges' totals were for the subjects.

Table 1 indicates that in only one case were the judges' ratings remarkably different; that rating is found in Criterion IV, Item 1, Rating A. This particular rating applies the criterion "knows when to draw an auxiliary segment" to a problem where no auxiliary segment is needed. The dissimilarity in the judges' ratings appeared to result from a discrepancy in how to rate an item when the criterion does not apply. One judge typically gave the benefit of the doubt and high ratings of 4; the other two used the "cannot be determined" rating of 1. Given this finding, the averages of the judges' scores were used as the student's score (unit of analysis on these two tests).

The typically high IRC's in Table 1 indicate the judges interpreted the scoring criteria and student performances similarly. It also means that total scores tend to represent the same pattern of performance across the elements of a proof and not different patterns for each judge. Total scores therefore are highly meaningful and interpretable and mean the same thing for different students who have the same score. Consequently, these results are evidence for both the reliability and validity of the proof scores obtained via the analytical rating method, and for the analytical rating scale producing, in general, similar scores for different judges.

Internal Consistency

Internal consistency refers to the interrelatedness of a series of items or scale. Items that focus on the same skill or trait tend to correlate with each other. Items that are not representative of the same latent trait tend to be less correlated with each other. Items aimed at measuring the same latent trait should be highly correlated to be reliable and valid. Therefore, internal consistency is tied to the correlation among items and is an index of coherence and

interpretability. It is in part a *reliability* index and in part a *validity* index for several different reasons (see below for details).

Measures of internal consistency assess both the interpretability and the validity of an instrument. Interpretability in the sense that subjects with the same total score have similar patterns of correct items and validity in the sense that the items are measuring the same latent trait. Furthermore, according to Nunnally (1967, p. 210), when a test is administered to a group only once a good way to estimate reliability is through measures of internal consistency. Cronbach's alpha (α) or its dichotomous equivalent, Kuder-Richardson Formula 20 (KR20), are both very good measures of internal consistency.

Table 2 shows either the Cronbach alpha or the KR20 for the four instruments used to assess the performance of the analytic rating scale. The instruments were the Equilibrium in a Balance Test (EBT), Revised van Hiele Test (RVHT), CDASSG Proof Test (CPT), and the New Proof Test (NPT). The Cronbach alphas reported were those measured on the mean ratings of the three judges on each judgment. The EBT and RVHT will be discussed in greater detail below.

Table 2 also compares the three methods of scoring the two proof tests, CPT and NPT. On the top line, the line with the proof test's name, are the statistics for the aggregate scores using the analytic-scoring method. Each rating on the analytic score sheet is regarded as an item. The proof tests were also subdivided into tests for each criterion. These subtests results are presented on the line just below the aggregate test results. For a breakdown of the items on the "Five Criteria Subtests" refer to the sample scoresheet in Table 1. The items for the "Holistic" method and the "Dichotomous" method are the actual proofs on the tests, of which there were six.

As can be seen in Table 2, most of the instruments used for the

present research had internal consistency reliabilities over .50. Of special note, when the analytic scoring method was applied, the alpha for the proof tests were over .97, whereas the alpha for those instruments that are dichotomous or holistically scored are much lower. Notice also that the alphas for the five criteria subtests of the analytic scoring method are also over .90.

Table 2: *Reliability Coefficients for Equilibrium in a Balance (EBT), Revised van Hiele (RVHT), CDASSG Proof (CPT), and New Proof (NPT) Tests*

Instrument	N	Items	KR20 or Cronbach's α
Equilibrium in a Balance (EBT)	239	15	KR20's = .54
Three Stage Subtests		5, 5, 5	(-, .70, .48)
Revised Van Hiele Test (RVHT)	235	24	KR20's = .68
Three Level Subtests		8, 8, 8	.47, .55, .26
CDASSG Proof Test (CPT)	122	60	α's = .99
Five Criteria Subtests		12, 12, 12, 9, 15	.95, .94, .96, .94, .97
Holistic Rating Scale		6	.89
Dichotomous Scale		6	.86
New Proof Test (NPT)	119	60	α's = .98
Five Criteria Subtests		12, 12, 12, 9, 15	.92, .94, .93, .91, .95
Holistic Rating Scale		6	.84
Dichotomous Scale		6	.77

Notes. The source of the reliabilities in parentheses is Adi and Pulos, 1978. The statistics on the top line of each proof test are for the analytic scoring method.

The very high alphas, seen in Table 2, indicate the scores from the proof tests are highly interpretable and very internally consistent across criteria scoring. We also see, considering that the proof tests were only administered once, that scoring the proof tests using the analytic scoring method results in a more reliable score. Thus, we conclude from Table 2 that using the analytic scoring method improves the internal consistency of the proof test, which in turn makes the proof test more reliable, interpretable, and valid.

The Distribution of Scores

We next considered the possible effects of the analytic scoring method on various analyses researchers tend to do. An expected effect of the method was a change in the distribution of scores. Descriptive statistics presented in Table 3 consist of *ranges, means* (M), *standard* deviations (SD), *skewness, and* Ikurtosis (Kurt) for the proof tests, CPT and NPT, by method of scoring. The methods of scoring are dichotomous (i.e., right or wrong), holistic, and analytic. The comparative statistics presented in Table 3 consist of one-way analysis of variance *F* ratios, *omega* (ω) squares, and Kolmogorov-Smirnov (K-Z) two-sample *Z* scores, which give a direct comparison of the tests' distributions of scores.

We observed that the analytic method of scoring generated more normal distributions for both the CDASSG Proof Test (CPT) and the New Proof Test (NPT). The descriptive statistics for comparing the NPT and CPT are given in Table 3. For example, we found the analytic method diminished the kurtosis of the NPT from -1.09 using the dichotomous method to -.64. We also found the skewness of the CPT decreased from .36 for the dichotomous ratings to .19. However, the analytic scoring method did not change the distribution of the NPT nor the CPT more than the holistic; in fact the holistic method had the best overall descriptive statistics. The holistic method shows better descriptive statistics because it is not as fine-grained a scale as the analytic rating scale and therefore skewness and kurtosis would need to be much more pronounced to affect the skewness and kurtosis indices. The same is true for a dichotomous scale. Bearing this in mind, it could be argued that the statistics for the analytic rating scale are better than either of the other methods.

Overall, the analytic method of scoring appears preferable when

correlational studies and analysis of variance are intended because normal distributions are assumptions for these types of analyses. The comparative statistics for the CPT and NPT using the three types of scoring methods are given in Table 3.

Table 3: *Descriptive and Comparative Statistics, by Various Scoring Methods, for the CDASSG Proof Test (CPT) and the New Proof Test (NPT) (N's = 122 CPT, 119 NPT)*

Scoring Method	Range	M	SD	Skew-ness	Kurt.	*F*	ω^2	*K-S Z*
Dichotomous						3.91 $p < .05$	.01	1.48 $p < .03$
CPT	6.0	2.5	2.2	.36	-1.35			
NPT	6.0	3.1	1.9	.13	-1.09			
Holistic						4.18 $p < .05$	.01	1.36 $p < .05$
CPT	21.3	12.9	5.9	.19	-1.23			
NPT	22.3	14.3	5.3	-.05	-.78			
Analytical						6.31 $p < .02$	.02	1.67 $p < .01$
CPT	179.7	129.7	52.1	.19	-1.24			
NPT	183.7	145.4	44.3	-.32	-.64			

The *F* ratios, omega squares, and K-S *Z*'s shown in Table 3 indicate the definite affects of using an analytic scoring method. Note how *F* gets larger and becomes more significant when the analytic method is used (*Fs* = 3.91, 4.18, 6.31; *p's < .05, .05, .02; for dichotomous, holistic, and analytic, respectively)*. Note how the amount of variance explained by the proof test form doubles when the analytic scoring method is used (ω^2.= .01, .01, .02 for dichotomous, holistic, and analytical, respectively). The *F* ratios and omega squares indicate the greater sensitivity of the analytic scoring method to the subtle differences between the test forms. Lastly, note that the using a more

sensitive scoring method can have a pronounced effect on a test's distribution of scores as indicated by the Kolmogorov-Smirnov *z* scores (*Z*'s = 1.48, 1.36, 1.67; *p*'s = .03, .05, .01; for dichotomous, holistic, and analytic, respectively).

We also investigated the hypothesized sensitivity of the analytic method of scoring by doing a step-down analysis of variance on the CPT and NPT using the methods of scoring for the steps. Table 4 presents the results of a step-down ANOVA of CPT and NPT controlling the covariance of the dichotomous method of scoring with the other methods of scoring.

Table 4: *Step-Down Analysis of Variance for CPT versus NPT Using Method of Scoring for Steps and Controlling for Dichotomous Scoring*

Scoring Method	Step-down *F*	Hypoth. *df*	Error *df*	Sig. of *F*
Dichotomous	3.91	1	239	.049
Holistic	.26	1	238	.609
Analytic	7.65	1	237	.006

Table 4, as an extension of Table 3, indicates another advantage of the analytic scoring method. While the *F* ratio for the holistic scoring method loses significance when covariance of the dichotomous test is controlled, as can be seen in Tables 3 and 4 (*p* increases from .049 to .609), the significance of the *F* ratio of the analytic scoring method becomes greater (*p* decreases from .02 to .006). The implication of the step-down ANOVA is that when the covariance is controlled, the holistic method of scoring (0 to 4 rating) adds no more information than the dichotomous right-or-wrong scoring method, while the analytic scoring method continues to yield a great deal of unique information about the nature of the tests and student performances. Clearly, the analytic scoring method can yield more information about the subjects or

instruments. Tables 3 and 4 show that the method of scoring can affect ANOVA results in several ways.

Next, we consider the effect method of scoring has on correlations. To consider correlations, the proof tests must be compared with other instruments. We will show the comparisons with two other instruments, the Revised Van Hiele Test (RVHT) and the Equilibrium in a Balance Test (EBT). Since this research utilized a convergent-discriminant design, the comparison of correlations was especially important. First we will elaborate the reasons these tests were used in the larger convergent-discriminant design.

The objective of the RVHT is to categorize students by van Hiele level, of which there are presently three. As such, the RVHT is really a combination of tests intended to sample from several domains. The items of the RVHT were not intended to sample a unique content area or skill, but to sample from three or more descriptive areas. As shown in Table 2, the RVHT had three subtests, one for each theorized van Hiele level, *visual*, *descriptive*, and, *theoretical*. The RVHT was scored two ways. Subjects received a simple total correct score for the entire 24 item test, RVHT; and they were assigned to a van Hiele level, RVHLEV, by the algorithm outlined above.

The EBT was a test designed to sort subjects into the cognitive stages theorized by Piaget. Only three of Piaget's theorized stages were under consideration, *concrete*, *transitional*, and *formal*. Thus, the EBT was divided into three subtests (see Table 2). The EBT was also scored in two ways, similar to the way the RVHT was scored. The scores associated with the EBT were EBT, for total correct, and EBLEV, for level assigned.

Table 5 presents examples of contrasting correlations for the two proof tests, CPT and NPT, using different methods of scoring, with the

RVHT and the EBT. The methods of scoring the proof tests are presented as before, using the proof variable name. Pearson product-moment correlations were used throughout this study. As can be seen from Table 5, correlations can vary depending on the scoring method that is used. The two proof instruments, CPT and NPT, vary to the extent they correlate with the EBT and RVHT both when the total scores are used (RVHT and EBT) and when the instruments' subtests are used to separate the subjects into level categories (RVHLEV and EBLEV).

Table 5: *Contrasting Correlations, Related to Method of Scoring, for CPT and NPT Samples with RVHLEV, RVHT, EBLEV, and EBLEV (N = 241)*

	CDASSG Proof Test (CPT) (N = 122)			New Proof Test (NPT) (N = 119)		
	Dichot.	Holistic	Analytic	Dichot.	Holistic	Analytic
RVHLEV	.46	.47	.45	.52	.52	.49
RVHT	.49	.52	.50	.57	.58	.54
EBLEV	.42	.40	.40	.46	.44	.42
EBT	.40	.39	.38	.45	.43	.40

Note. $p < .01$ for all values in table.

Though there is no significant difference in the correlations on any row for either the CPT or the NPT, Table 5 indicates the analytic scoring method correlations are usually the lowest correlations in a row, particularly for the NPT. As Ferguson notes (1981, p. 107), correlation estimates the magnitude of concomitant variation. Hence, the nature of the distributions of the scores to be correlated affects the size of the correlation coefficient. The more normal the standardized distributions of the variables, the more likely the correlation coefficient observed will be untruncated.

The analytic scoring method resulted in more normal distributions of scores (K-S $Z's$ = .57, 1.05; $p's < .90, .22$, for the NPT and CPT,

respectively), than the RVHLEV (K-S z = 3.25, $p < .001$), EBT (K-S z = 1.56, $p < .02$), and EBLEV (K-S z = 4.84, $p < .001$), and than those of the other two scoring methods, dichotomous (K-S $Z's$ = 1.44, 2.09; $p's <$ .03, .001, for the NPT and CPT, respectively), and holistic (K-S $Z's$ = .87, .96; $p's <$ .44, .31, for the NPT and CPT, respectively). Hence, in general, the correlations for the analytic scoring method are more accurate than those for the holistic and dichotomous methods as estimates of the population parameters for the coefficients.

Validity of the Subtests

We viewed the analytic scoring criteria as being fairly hierarchical. As such, we considered using the ratings for each of the five criteria as subtests that could be used to sort or categorize the subjects by levels of proof mastery or performance (unschooled to expert as outlined above). Consequently, we decided to test the hierarchy of the criteria (see Table 6 for criteria).

As previously mentioned, a method of estimating the reliability of an instrument to hierarchically categorize subjects is the Guttman Scalogram method (McIver & Carmines, 1981, p. 40). The Guttman Scalogram method is an analysis of the subjects' subtest score patterns. Each subtest is scored as either a success or failure. If a 1 means a success and a 0 means a failure, then a pattern consists of a series of 1's and 0's, written from left to right, indicating the pass or fail status from the lowest category subtest to the highest. The scalogram method estimates how reproducible and scaleable the subject patterns are.

To have a *scaleable* pattern the subject must succeed at all levels (e.g., categories) before the highest level mastered. Patterns such as (1, 0, 0) and (1, 1, 0) are scaleable patterns because, starting at the left and reading to the right, the first 0 encountered is followed by an

uninterrupted series of 0's. Patterns such as (0, 1, 0) and (1, 0, 1) are not scaleable. To be *reproducible* means that the pattern of successes can be determined from the total score. For example, a total score of 3 on a 5 category test must imply that the subject was successful on the first three categories of the five.

Calculating the Guttman reproducibility coefficient involves determining the number of corrections that must be made to make the patterns of all subjects scaleable and reproducible, dividing that number by the number of possible changes that could have been made. This quotient is then subtracted from 1 to determine the reproducibility of the subtests. The calculations result in what is known as a coefficient of reproducibility. Reproducibility coefficients on or above .90 are regarded as significant (McIver & Carmines, 1981, p. 51).

In addition to Guttman's coefficient of reproducibility, there is also a coefficient of scalability. The coefficient of scalability is calculated the same way as the coefficient of reproducibility except for the counting of errors. For the coefficient of scalability the subject's total score is considered inviolate and corrections are made so that the response pattern is scaleable for that total score. This is usually a larger number of corrections and results in a smaller coefficient.

The coefficient of scalability is considered an indicator of possible improvement in prediction when the scale is used as opposed to using the total-correct scores. A coefficient of .60 or more is considered significant for scalability (McIver & Carmines, 1981, p. 51).

Using cutoffs of 75% of total possible ratings to establish success for a criterion, we calculated the Guttman reproducibility (CR) and scalability (CS) coefficients for both proof tests. The NPT and CPT reproducibility coefficients were .86 and .95, respectively; the

scalability coefficients were .72, and .81. These numbers indicate that the analytic scoring criteria are scaleable and that the there is some benefit to using the scale generated by the criteria. Interpreting the five scoring criteria as roughly corresponding to the student behavior categories, unschooled to expert, the scalability of the analytic scoring criteria provide a new way of categorizing students which we will present below.

Since the analytic scoring criteria were not uniformly applied to all six proofs on a test, we considered the reproducibility coefficients of the analytic scoring criteria for particular proofs. All the proofs on both tests had significantly high reproducibility coefficients. The coefficients ranged from .96 to 1.00 for the NPT and .92 to 1.00 for the CPT. In particular, using the analytic scoring method, Proof 1 on the NPT was rated for Criteria III (knows relevant propositions), IV (has a plan), and V (succeeds). The reproducibility and scalability coefficients for the NPT's Proof 1 were both equal to 1.00 because the subject patterns were all reproducible and scaleable.

Given the evidence above, we concluded that the analytic scoring method was very scaleable on all criteria and aggregate scores, and that we could use the ratings for each criterion as subtests. Such evidence of scalability also supports the validity of the criteria as measurable behaviors. We, therefore, believe the analytic scoring method yields subtest scores that are highly descriptive of the subject's performance; in fact, much more so than the other two scoring methods.

Comparing proof scores to assigned revised van Hiele level

As previously mentioned, the Revised van Hiele Test was developed to sort students by hierarchical categories of geometric thinking. The van Hiele level instrument (RVHT), consisted of three subtests of eight items apiece. According to the theory, the van Hiele levels are

hierarchical beginning with *visual-recognition,* then *descriptive,* and thirdly *theoretical.* Table 6 shows correlations for the analytic scoring criteria with the three level subtests of the RVHT. These correlations are shown for both the NPT and the CPT.

The correlations presented in Table 6 also suggest that the RVHT Level 3 subtest is not the best predictor of proof performance of the three RVHT subtests. Our results indicated that the RVHT is also highly scaleable and thus hierarchical (CR = .97, CS = .80). However, one sees in Table 6 that generally the highest correlations are found in the Criterion IV row. This finding suggests that the "having a plan" criterion best predicts performance on any of the RVHT level subtests. As having a plan is classically considered to be a prime exemplar of higher order cognitive activity, this result strongly supports the model of doing proofs we have outlined above and the validity of the analytic rating scale devised.

Table 6: *Correlations by Analytic Scoring Standards applied to the CDASSG Proof Test and the New Proof Test with the Subtests of the Revised Van Hiele Test*

	CDASSG Proof Test (CPT)			New Proof Test (NPT)		
	Visual	Descrp.	Theor.	Visual	Descrp.	Theor.
Criterion I, knows terms, marks, notation	.36	.40	.31	.42	.42	.29
Criterion II, comprehends problem	.36	.37	.31	.40	.37	.27
Criterion III, knows relevant propositions	.33	.38	.29	.38	.44	.32
Criterion IV, has a plan/draws needed parts	.37	.39	.33	.44	.46	.34
Criterion V, succeeds at proof or problem	.34	.35	.31	.37	.39	.35

Note. $p < .01$ for all correlations.

Finally, Table 6 also shows that the van Hiele theory does not align itself with the integration of Bloom's taxonomy and Greeno's hypothesized areas of geometric knowledge presented earlier. Recall

that, the scoring criteria we used were based on a combination of Bloom's and Greeno's theories. If both the van Hiele theory and the analytic scoring criteria represented the same hierarchical construct, the correlations would be largest for Criteria IV and V with RVHT Level 3, theoretical. They are not. Table 6 shows that, in general, the highest correlations occur at Level 2 (descriptive). One explanation of this result might be that the "Y" structure of Bloom's taxonomy confounds its alignment and agreement with van Hiele's theory. Another explanation might be that the analytic scoring criteria and the revised van Hiele theory do not represent the same construct. We believe that the latter explanation is the more correct one.

Performance Classifications

We also found that the analytic scoring criteria could be used to classify subjects into the broad performance categories discussed earlier. We used the following method for classification:

1. Use 75% of the possible total score on a criterion as the success cut-off. Score the criterion as a one if cut-off is reached, otherwise score as a zero.

2. Classify (0,0,0,0,0) patterns as "Unschooled."

3. Classify (1,0,0,0,0) patterns as "Novices."

4. Classify (1,1,0,0,0) or (1,1,1,0,0) patterns as "Intermediates."

5. Classify (1,1,1,1,0) patterns as "Competents."

6. Classify (1,1,1,1,1) patterns as "Experts."

Of the 241 students who took a proof test, 185 (77%) were classified into one of the performance categories and the classifications were strongly correlated with the proof scores ($r = .86$, $p < .001$). Of those that were classified, 67% were unschooled, 4% were Novices, 4% were Intermediates, 7% were Competents,

and 18% were Experts. This finding indicates that only 25% of those classified, or 46 of all 241 subjects (19%) were capable of proof at the end of a year's study of geometry. This finding was not unexpected as previous research also found only a small percentage of geometry students had mastered geometric proof (e. g., Senk, 1985, 30%).

These findings are evidence of both the strength and validity of the classification system based on the analytic scoring criteria and the criteria themselves. When subject performance classifications (e.g., Novice) were correlated with van Hiele level assignments, the was only a moderate correlation ($r = .53$), which is respectable as there is no chance of mapping three categories unto five perfectly.

This finding is evidence that the van Hiele level assigned a subject is not as strong a predictor of proof performance as one would like to see or as suggested by advocates of the van Hiele model for a valid model of geometry performance. The moderate correlation for the revised van Hiele level assigned with the proof performance classifications coupled with the correlations shown in Table 6 lead to an unexpected conclusion. The unexpected conclusion is: Either the Revised Van Hiele Test (RVHT) is not a strong predictor of proof performance or students with the geometric thinking level that would enable them to do proof are not mastering the skill. This conclusion suggests that either the RVHT be corrected to accurately predict proof performance, or instruction is not enabling students with the ability to accomplish proofs successfully, or the revised van Hiele theory is not yet an adequate model for predicting proof performance even if it is corrected. Given the weight of the research evidence presented in this study, we believe that the van Hiele test needs to be revised again, and, due to the limited descriptions of the levels, the van Hiele theory is not yet an adequate model for predicting proof performance.

Conclusions

This research was intended to investigate three methods of scoring geometry proofs. The three methods were dichotomous (right-or-wrong), holistic, and analytical. To do this comparison, a new analytical method of scoring proofs was developed. The genesis of the analytic method based on the theories of Greeno and Bloom was discussed.

Tables 1 and 2 have indicated at least two benefits of the analytic scoring method. First, the method results in very similar patterns of ratings for judges indicating a very reliable set of criteria that give dependable results. Second, the method produces higher levels of internal consistency, which indicates that these scores are more interpretable, and thus more valid.

Tables 3 and 4 have suggested two benefits of the analytic method of scoring. The analytic method results in a better distribution of scores through increasing the range of values, and the increased range improves the interpretability and descriptive capacity of the scores themselves. Secondly, the analytic scoring increases the likelihood of significant, reliable, and interpretable findings in ANOVA and in correlational studies. This latter benefit is greater accuracy and sensitivity in hypothesis testing.

Tables 5 and 6 also showed that there was another benefit of the analytic scoring method. The ratings for each criterion could be used to form highly scaleable and reproducible subtests. This result reinforces our view that the analytic scoring method is more descriptive of the subjects' performance. We also showed that the subtests can be used in correlational studies revealing similarities and differences with other performance theories. In particular, we have seen that the van Hiele theory is not aligned with Bloom's when Greeno's hypothesized knowledge areas are integrated into the taxonomy.

Further, the analytic scoring criteria could be used to sort

roughly 77% of subjects into proof performance categories ranging from unschooled to expert. Of those classified, by far the highest percentage of students fell into the most unskilled class, the unschooled (67%). After a year of geometry, only 25% of those classified could be considered competent or expert at geometric proof. The remaining 8%, those classified as Novices or Intermediates, were possibly just acquiring necessary proof skills, which are considered integral parts of formal geometry study.

All of these results support four conclusions. These four conclusions are:

1. The analytic scoring method results in better internal consistency and reliability indices than the dichotomous or holistic approaches. The increased number of scale points directly affects the internal consistency of the tests. The high level of interrater reliability also supports the view that a coherent and dependable set of criteria was developed.

2. The increased number of scale points also improves the distribution of scores. The larger range of scores results in more interpretable scores and a more normal distribution. A more normal distribution of scores affects subsequent analyses such as ANOVA.

3. The scoring criteria make possible a better interpretation of a subject's performance. The criteria were shown to be highly scaleable and reproducible suggesting a strongly hierarchical structure. Therefore, subject performance on criteria subtests is highly interpretable and aids in more specific descriptions of student proof performance.

4. Lastly, the analytic scoring criteria themselves can be used as both independent and dependent variables in analyses. We saw the criteria could be used to contrast Blooms's taxonomy with the van Hiele theory. The scoring criteria could also be used to explore construct validity for other theories related to proof, and the scoring criteria could be used to classify students into categories of proof performance.

In light of these demonstrable benefits, we highly recommend the use of this analytic scoring method over holistic or dichotomous methods. We believe it to be a method that can be used by researchers and classroom teachers alike. The benefits of use by classroom teachers would include better feedback to students, more specific information to guide

curriculum development for teachers (and researchers), and a method for diagnosing learning inequities related to proof. We also believe that the model of the factors and processes involved in doing proofs that we developed utilizing the analytic rating scale presented here is a better model than van Hiele's, Greeno's or others currently in the literature.

References

Adi, H. (1976). *The interaction between the intellectual development levels of college students and their performance on equation solving when different reversible processes are applied.* Unpublished doctoral dissertation, Florida State University.

Adi, H. (1980). Individual differences and formal operational performance of college students. *Journal for Research in Mathematics Education, 11*, 150-156.

Bloom, B. S. (Ed.). (1956). *Taxonomy of educational objectives.* New York, NY: David McKay Company, Inc.

Charles, R., Lester, F., & O'Daffer, P. (1987). *How to evaluate progress in problem solving.* Reston, VA: National Council of Teachers of Mathematics.

Campbell, D. T., Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81-105.

Ferguson, G. A. (1981). *Statistical analysis in psychology and education* (5th ed.). New York, NY: McGraw-Hill.

Glaser, R. (1978). *Advances in instructional psychology.* Hillsdale, NJ: Lawrence Erlbaum.

Glover, J. A., Ronning, R. R., & Bruning, R. H. (1990). *Cognitive psychology for teachers.* New York, NY: Macmillan.

Greeno, J. G. (1978). A study in problem solving. In R. Glaser (Ed.).

Advances in instructional psychology: Vol. 1 (pp. 13-75). Hillsdale, NJ: Lawrence Erlbaum Associates.

Kerlinger, F. N. (19730. *Foundations of behavioral research* (3rd ed.). Philadelphia, PA: Holt, Rinehart, and Winston, Inc.

Lester, F. K. (1980). Research on mathematical problem solving. In R. W. Shumway (Ed.). *Research in mathematics education* (pp. 286-323). Reston, VA: National Council of Teachers of Mathematics.

McIver, J. P., & Carmines, E. G. (1981). *Unidimensional scaling.* Beverly Hills, CA: Sage Publications.

Newell, A., & Simon, H. A. (1972). Human problem solving. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Nunnally, J. C. (1967). *Psychometric theory.* New York, NY: McGraw-Hill.

Polya, G. (1957). *How to solve it* (2nd ed.). New York, NY: Doubleday.

Teppo, A. (1991). Van Hiele levels of geometric thought revisited. *The Mathematics Teacher, 84,* 210-221.

402

Test Number _____

REVISED VAN HIELE TEST

Directions

Do not open this test booklet until you are told to do so.

This test contains 24 geometry questions and ten problem-solving questions. It is not expected that you know everything on the test.

There is a test number on the top right hand corner of this page. Write this number where you are instructed on your answer sheet.

When you are told to begin:

1. Read each question carefully.
2. Decide upon the answer you think is correct. There is only one correct answer to each question. Fill in the circle corresponding to your answer on the answer sheet.
3. Use the scrap paper for figuring or drawing. **DO NOT MARK THE TEST.**
4. To change an answer, completely erase your previous answer.
5. If you need another pencil raise your hand.
6. You will have 40 minutes for this test.

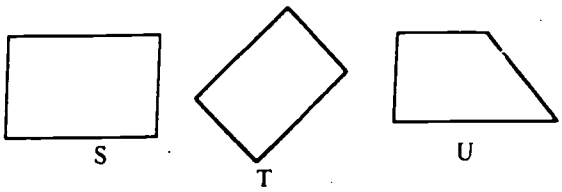
Wait until your teacher says that you may begin.

Figure E.2. Cover page for combination of RVHT and PROBSOLV.

Revised Van Hiele Test

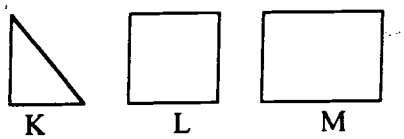
1. Choose the rectangle or rectangles.

(A) S only
(B) T only
(C) S and T only
(D) S and U only
(E) All are rectangles.



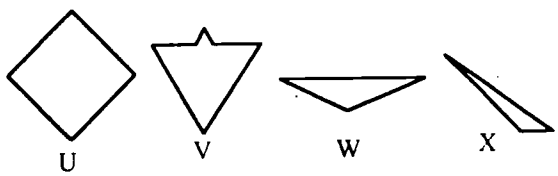
2. Choose the square or squares.

(A) K only
(B) L only
(C) M only
(D) L and M only
(E) All are squares.



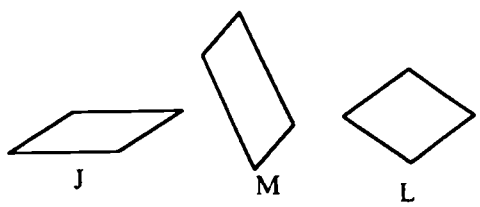
3. Choose the triangle or triangles.

(A) None of these are triangles.
(B) V only
(C) W only
(D) W and X only
(E) V and W only



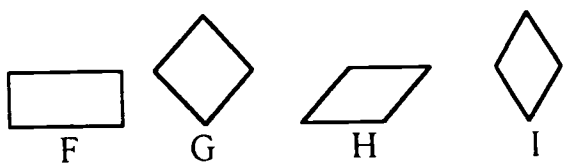
4. Choose the parallelogram or parallelograms.

(A) J only
(B) L only
(C) J and M only
(D) None of these are parallelograms.
(E) All are parallelograms.



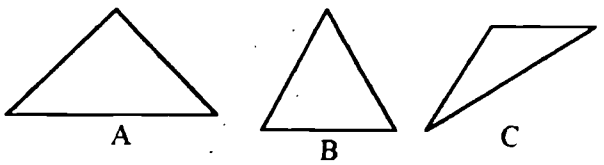
5. Choose the square or squares.

(A) None of these are squares.
(B) G only
(C) F and G only
(D) G and I only
(E) All are squares.



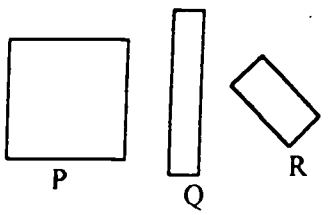
6. Choose the shape or shapes that can be called right triangles.

(A) A only
(B) B only
(C) C only
(D) A and B only
(E) None of (A)—(D)



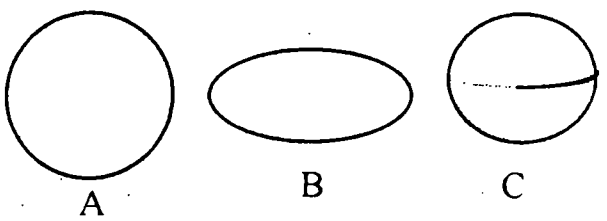
7. Which of these can be called a rectangle.

(A) All can.
(B) Q only
(C) R only
(D) P and Q only
(E) Q and R only



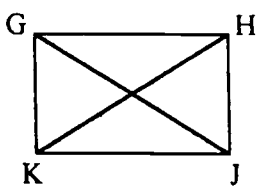
8. Choose the circle or circles.

(A) A only
(B) A and C only
(C) C only
(D) B only
(E) All are circles.

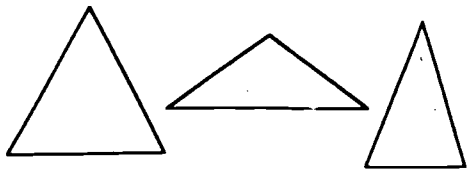


9. In a rectangle GHJK, $\overline{GJ}$ and $\overline{HK}$ are the diagonals.

Which of (A)—(D) is **not** true in **every** rectangle?
(A) There are four right angles.
(B) There are four sides.
(C) The diagonals have the same length.
(D) The opposite sides have the same length.
(E) All of (A)—(D) are true in every rectangle.



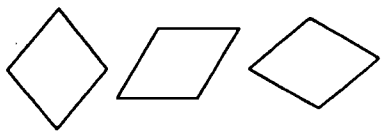
10. An isosceles triangle is a triangle with two sides of equal length.



Here are three examples.
Which of (A)—(D) is true in every isosceles triangle?
(A) The three sides must have the same length.
(B) One side must have twice the length of another side.
(C) There must be at least two angles with the same measure.
(D) The three angles must have the same measure.
(E) All of (A)—(D) are true in every isosceles triangle.

11. A rhombus is a 4-sided figure with all sides of the same length.

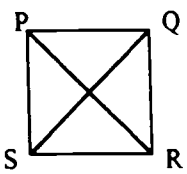
Here are three examples.



Which of (A)—(D) is not true in every rhombus?
(A) The two diagonals have the same length.
(B) Each diagonal bisects two angles of the rhombus.
(C) The two diagonals are perpendicular.
(D) The opposite angles have the same measure.
(E) All of (A)—(D) are true in every rhombus.

12. What characteristic do all rectangles have that some parallelograms do not have?
(A) opposite sides equal
(B) diagonals equal
(C) opposite sides parallel
(D) opposite angles equal
(E) None of (A)—(D)

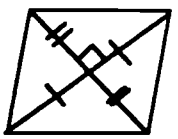
13. PQRS is a square.
Choose the relationship that is true of all squares.



(A) $\overline{PR}$ and $\overline{RS}$ have the same length.
(B) $\overline{QS}$ and $\overline{PR}$ are perpendicular.
(C) $\overline{PS}$ and $\overline{QR}$ are perpendicular.
(D) $\overline{PS}$ and $\overline{QS}$ have the same length.
(E) Angle PQR is larger than angle SRQ.

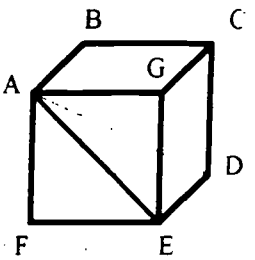
14. Classify the following shape using the markings. It may not be drawn accurately. Give it the **most specific** name possible. [Hint: The same number of notches means the lengths are the same, the box at the center means it is a right angle.]

(A) polygon.
(B) quadrilateral.
(C) rectangle.
(D) rhombus
(E) square



15. The figure below is a cube. Which relationship is true of all cubes?

A) $\overline{AD}$ and $\overline{FE}$ are the same length.
B) $\overline{AF}$ is parallel to $\overline{CD}$.
C) $\overline{BC}$ and $\overline{AE}$ are the same length.
D) $\overline{AD}$ and $\overline{AE}$ are the same length.
E) Angle ABC is greater than angle AFE.



16. What sentence **best** describes a square?

(A) It is a rectangle with all sides the same length.
(B) It is a 4-sided figure with sides that are the same length.
(C) It is a quadrilateral with four right angles.
(D) It is a 4-sided figure with all sides the same length and the opposite sides are parallel.
(E) It has four right angles and the opposite sides are parallel.

17. Here are two statements.

Statement S: ΔABC has three sides of the same length.
Statement T: In ΔABC, $\angle B$ and $\angle C$ have the same measure.

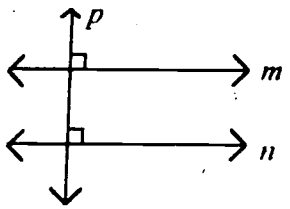
Which is correct?

(A) Statements S and T cannot both be true.
(B) If S is true, then T is true.
(C) If T is true, then S is true.
(D) If S is false, then T is false.
(E) None of (A)—(D) is correct.

18. Examine these three sentences.

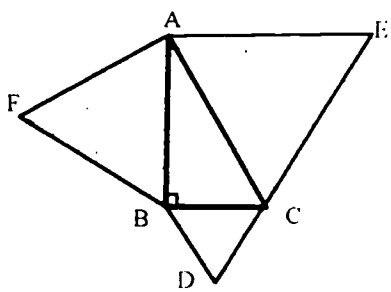
(1) Two lines perpendicular to the same line are parallel.
(2) A line that is perpendicular to one of two parallel lines is perpendicular to the other.
(3) If two lines are equidistant, then they are parallel.

In the figure below, it is given that lines m and p are perpendicular and lines n and p are perpendicular. Which of the sentences above could be the reason that line m is parallel to line n?



(A) (1) only
(B) (2) only
(C) (3) only
(D) Either (1) or (2)
(E) Either (2) or (3)

19. Here is a **right** triangle, ΔABC. Equilateral triangles, ΔACE, ΔABF, and ΔBCD, have been drawn on the sides of ΔABC.



From this information, one can prove that $\overline{AD}, \overline{BE}$, and $\overline{CF}$ have a point in common. What would this proof tell you?

(A) Only for ΔABC drawn above can we be sure $\overline{AD}, \overline{BE}$, and $\overline{CF}$ have a point in common.
(B) In some but not all right triangles, $\overline{AD}, \overline{BE}$, and $\overline{CF}$ have a point in common.
(C) In any right triangle, $\overline{AD}, \overline{BE}$, and $\overline{CF}$ have a point in common.
(D) In any triangle, $\overline{AD}, \overline{BE}$, and $\overline{CF}$ have a point in common.
(E) In any equilateral triangle, $\overline{AD}, \overline{BE}$, and $\overline{CF}$ have a point in common.

20. Here are two statements.

I. If a figure is a rectangle, then its diagonals bisect each other.
II. If the diagonals of a figure bisect each other, then the figure is a rectangle.

Which of the following statements is (are) correct?

(A) To prove I is true, it is enough to prove two is true.
(B) To prove II is true, it is enough to prove I is true.
(C) To prove II is true, it is enough to find one rectangle whose diagonals bisect each other.
(D) To prove II is **false**, it is enough to find one non-rectangle whose diagonals bisect each other.
(E) None of (A)—(D) is correct.

21. In geometry:

(A) Every term can be defined and every true statement can be proved true
(B) Every term can be defined but it is necessary to assume certain statements are true.
(C) Some terms must be left undefined but every true statement can be proved true.
(D) Some terms must be left undefined and it is necessary that certain statements be assumed.
(E) None of (A)—(D) is correct.

22. Here are two statements.

Statement 1: The radii of circle A are the same length as the radii of circle B.
Statement 2: Circle A is the same size as circle B.
Which is (are) correct?

(A) If statement 1 is true, then statement 2 is false.
(B) If 2 is true, then statement 1 is false.
(C) Statements 1 and 2 cannot both be false.
(D) Either statements 1 and 2 are both true or both false.
(E) None of A—D is correct.

23. It is given that the measures of ∠A (angle A) and ∠B total 90° and that the measures of ∠C and ∠B total 90°. Based on the given, choose the statement or statements below that are valid conclusions?

I. ∠A and ∠B are the same size.
II. ∠A and ∠C total 90°.
III. ∠A and ∠C are the same size.

(A) I only
(B) I and II only
(C) II only
(D) II and III only
(E) III only

24. The lines *l*, *m*, and *n* **may <u>not</u>** be in the same plane. However, $l \perp m$ at point A, $n \perp m$ at point B, and A and B are not the same point. Which of the following conclusions would be valid?

(A) $l \perp n$.
(B) l (**not** $\perp$) n
(C) It cannot be determined if l is $\perp$ to n.
(D) l is // to n.
(E) l and n will cross eventually.

ERB Problem Solving

25. To raise money a class made 100 plant hangers from twine costing $30.00. Other expenses totaled $15.00. If all of the hangers were sold for $1.25 each, what was the total profit?

(A) $95.00
(B) $80.00
(C) $70.00
(D) $45.00

26. A bowl of punch has exactly three ingredients, X, Y, and Z, in the proportion 1:2:6, respectively. What fractional part of the punch is X?

(A) $\frac{1}{9}$

(C) $\frac{1}{6}$

(B) $\frac{1}{8}$

(D) $\frac{1}{2}$

27. A certain utility company charges $15 for the first 100 cubic meters of gas consumed and $0.07 for each additional cubic meter consumed. What is the total charge for 2,800 cubic meters of gas?

(A) $174

(C) $196

(B) $189

(D) $204

28. On the average, Joe's car uses 2 liters of gasoline for every twenty-one kilometers of travel. How many kilometers does the car travel, on the average, on a full tank that holds 50 liters of gasoline?

(A) 2,100.0

(C) 105.0

(B) 525.0

(D) 52.5

29. Mr. Johnson owes a store $200 plus 1½ percent interest. What is his total debt to the store?

(A) $230.00

(C) $201.30

(B) $203.00

(D) $200.00

30. A quantity P is three less than twice Q. Which equation states this relationship?

(A) $P = \frac{2Q}{3}$

(C) P = 2Q - 3

(B) P=3 - 2Q

(D) P - 3 = 2Q

31. At a price of x grams for 1 dollar, how many grams can be bought for y dollars?

(A) $\frac{x}{y}$ (C) xy

(B) $\frac{y}{x}$ (D) y + x

32. If a rope is to reach from the top of a twelve meter pole to a point on the ground that is 16 meters from the base, then the minimum length of the rope must be how many meters?

(A) 32 (C) 24

(B) 28 (D) 20

33. A survey of television viewers showed that viewers were four times as likely to be watching program A as to be watching program B. If the chances that a randomly selected viewer will be watching program B are 1 out of 10, what are the chances that the viewer will be watching program A?

(A) 1 out of 40 (C) 1 out of 4

(B) 4 out of 40 (D) 4 out of 10.

34. The length of a rectangular floor is 1 meter more than twice the width of the floor. If the area of the floor is 36 square meters, then the length of the floor, in meters, is

(A) 4.5 (C) 9

(B) 6 (D) 18

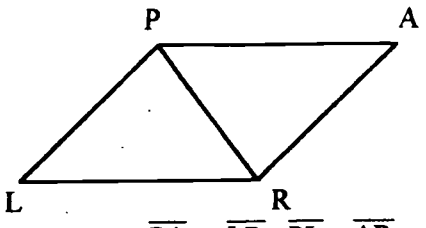
35. A shelf contains a total of 180 science, math, and English books. If there are twice as many English books as math books, and three times as many math books as science books, how many English books are on the shelf?

(A) 60 (C) 108

(B) 72 (D) 120

NRHS Proof Test

1. Write a Statement and Reason proof.



Given: $\overline{PA} \cong \overline{LR}$, $\overline{PL} \cong \overline{AR}$

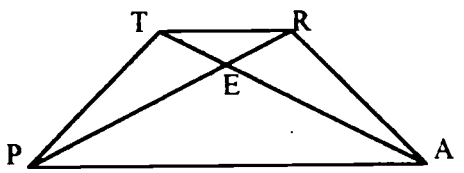
Prove: $\Delta PAR \cong \Delta RLP$

2. Explain why a triangle cannot have both an obtuse and a right angle. Use the indirect method of proof if you are familiar with it.

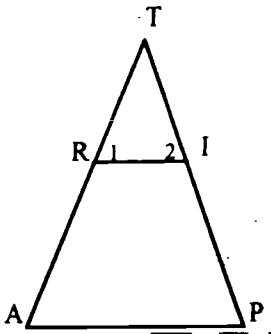
3. Write a Statement and Reason proof.

Given: TRAP is an isosceles trapezoid with $\overline{TP} \cong \overline{RA}$

Prove: $\overline{TA} \cong \overline{RP}$



4. Complete the Statement and Reason proof.



Given: $\overline{TR} \cong \overline{TI}$, $\overline{RI}$ // $\overline{AP}$

Prove: $\overline{RA} \cong \overline{IP}$

Statement	Reason
1. $\overline{TR} \cong \overline{TI}$, $\overline{RI}$ // $\overline{AP}$	1. Given
2. $\angle 1 \cong \angle 2$	2.
3. $\angle 1 \cong \angle A, \angle 2 \cong \angle P$	3.
4. $\angle A \cong \angle P$	4. Substitution
5. $\overline{TA} \cong \overline{TP}$	5.
6. $\overline{RA} \cong \overline{IP}$	6.s

5. Statement: **The diagonals of a rectangle are congruent.**

Suppose you wished to prove the above statement. In the space provided: 1. Draw and label a figure. 2. Write in terms of your figure, what is given and what is to be proved.

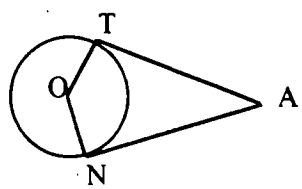
Figure:

Given:

Prove:

DO NOT PROVE THE STATEMENT ABOVE

6. Write a Statement and Reason proof. Prove the theorem: **Tangent segments from an external point to a circle are congruent**



Given: $\overline{TA}$ and $\overline{NA}$ are tangent to circle O

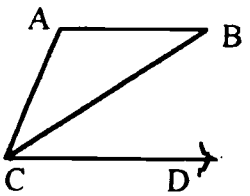
Prove: $\overline{TA} \cong \overline{NA}$

Extra: On the back, if time permits, explain why you were stumped by one of these proofs.

CDASSG Proof Test

1. Complete the Statement and Reason proof.

Statement	Reason
1. $\overline{AB}$ // $\overline{CD}$	1. Given
2. $\angle B \cong \angle BCD$	2.
3. AB = AC	3. Given
4.	4. Base angles of an isosceles triangle are congruent (equal in measure).
5.	5. Transitive property or substitution
6.	6. Definition of an angle bisector.



Given: $\overline{AB}$ // $\overline{CD}$,
AB = AC
Prove: $\overline{CB}$ bisects $\angle ACD$

2. Statement: **If a line passes through the midpoints of two sides of a triangle, it is parallel to the third side of the triangle.**

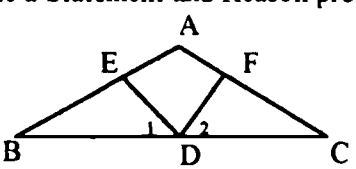
Suppose you wished to prove the above statement. In the space provided:
1. Draw and label a figure. 2. Write in terms of your figure, what is given and what is to be proved.

Figure: Given:

Prove:

DO NOT PROVE THE STATEMENT ABOVE

3. Write a Statement and Reason proof. Proof



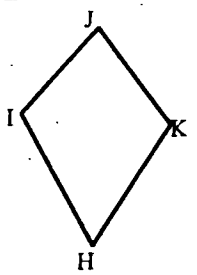
Given: D is the midpoint of $\overline{BC}$,
$\angle 1 \cong \angle 2$, $\overline{DE} \cong \overline{DF}$.
Prove: ΔABC is isosceles.

4. Write a Statement and Reason proof.

Given: Quadrilateral HIJK, HI = HK, IJ = JK

Prove: $\angle I \cong \angle K$

Proof



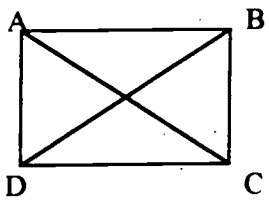
5. Write a Statement and Reason proof. Prove the theorem:

The diagonals of a rectangle are congruent.

Given: ABCD is rectangle.

Prove: $\overline{AC} \cong \overline{BD}$

Proof



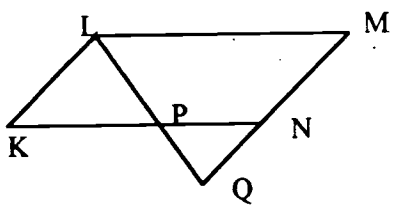
6. Write a Statement and Reason proof.

Given: KLMN is a parallelogram. N is on line $\overline{MQ}$

$\overline{LQ}$ and $\overline{KN}$ intersect at P.

Prove: $\Delta KLP \sim \Delta NQP$

Proof

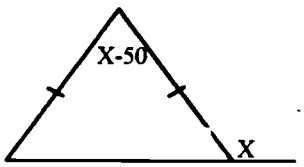


Extra: On the back, if time permits, explain why you were stumped by one of these proofs.

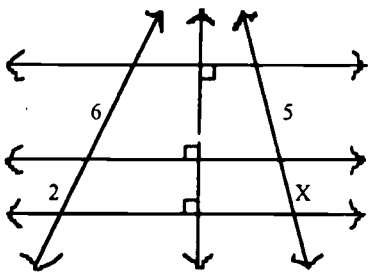
Problem-Solving Test

General instructions: Show all work. If you use a calculator, show how you set up the problem.

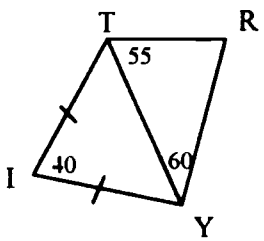
1. Solve for X.



2. Solve for X

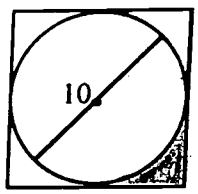


3. Name the **smallest** segment (side) in the diagram. Explain your choice.



The diagram above may be drawn inaccurately.

4. Find the area of the shaded region. The diameter of the circle is 10.



5. Triangle ABC has sides 3, 4, and 5. Triangle XYZ has sides 3, 4, and 6. Which triangle, ABC or XYZ, has the larger area? Explain your answer.

Extra: If you were stumped by a problem, explain what you believe you needed to know in order to reach a solution.

American Educational Research Association
April 18-22, 1995

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

TM025868

REPRODUCTION RELEASE

(Specific Document)

AERA /ERIC Acquisitions
The Catholic University of America
210 O'Boyle Hall
Washington, DC 20064

I. DOCUMENT IDENTIFICATION:

Title: Empirical Results of Using an Analytical versus Holistic Scoring Method to Score Geometric Proofs: Linking	
Author(s): Bethe McBride & James Ciarfio	
Corporate Source:	Publication Date: Feb. 1996

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic/optical media, and sold through the ERIC Document Reproduction Service (EDRS) or other ERIC vendors. Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce the identified document, please CHECK ONE of the following options and sign the release below.

☑ Sample sticker to be affixed to document — Sample sticker to be affixed to document ☒ disreg

Check here

Permitting microfiche (4" x 6" film), paper copy, electronic, and optical media reproduction

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Level 1

"PERMISSION TO REPRODUCE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Level 2

or here

Permitting reproduction in other than paper copy

Sign Here, Please

Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but neither box is checked, documents will be processed at Level 1.

"I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce this document as indicated above. Reproduction from the ERIC microfiche or electronic/optical media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries."

Signature: Bethe McBride	Position: Author
Printed Name: Bethe McBride	Organization: University of MA @ Lowell
Address: 42 Casablanca Ct. Haverhill, MA 01832	Telephone Number: (508) 373-0249
	Date: Feb. 23, 1996

You can send this form and your document to the ERIC Clearinghouse on Assessment and Evaluation. They will forward your materials to the appropriate ERIC Clearinghouse. ERIC/AERA Acquisitions, ERIC Clearinghouse on Assessment and Evaluation, 210 O'Boyle Hall, The Catholic University of America, Washington, DC 20064, (800) 464-3742

CUA

THE CATHOLIC UNIVERSITY OF AMERICA
Department of Education, O'Boyle Hall
Washington, DC 20064
202 319-5120

March 1995

Dear AERA Presenter,

Congratualations on being a presenter at AERA. The ERIC Clearinghouse on Assessment and Evaluation would like you to contribute to ERIC by providing us with a written copy of your presentation. Submitting your paper to ERIC ensures a wider audience by making it available to members of the education community who could not attend the session or this year's conference.

Abstracts of papers that are accepted by ERIC appear in RIE and are announced to over 5,000 organizations. The inclusion of your work makes it readily available to other researchers, provides a permanent archive, and enhances the quality of RIE. Your contribution will be accessible through the printed and electronic versions of RIE, through the microfiche collections that are housed at libraries around the country and the world, and through the ERIC Document Reproduction Service.

We are gathering all the papers from the AERA Conference. We will route your paper to the appropriate clearinghouse and you will be notified if your paper meets ERIC's criteria. Documents are reviewed for contribution to education, timeliness, relevance, methodology, effectiveness of presentation, and reproduction quality.

To disseminate your work through ERIC, you need to sign the reproduction release form on the back of this letter and include it with **two** copies of your paper. You can drop of the copies of your paper and reproduction release form at the ERIC booth (615) or mail to our attention at the address below. Please feel free to copy the form for future or additional submissions.

Mail to: AERA 1995/ERIC Acquisitions
The Catholic University of America
O'Boyle Hall, Room 210
Washington, DC 20064

Sincerely,

Lawrence M. Rudner, Ph.D.
Director, ERIC/AE